Welcome to

instats

The Session Will Begin Shortly

START

Statistics in R with Tidyverse

Session 7: Sampling

Statistics vs Parameters

- Definitions
  - Statistics: estimates from a sample
  - Parameters: true values in the population
- Importance of distinction
  - Accuracy and reliability of inferences
  - Application to real-world data analysis

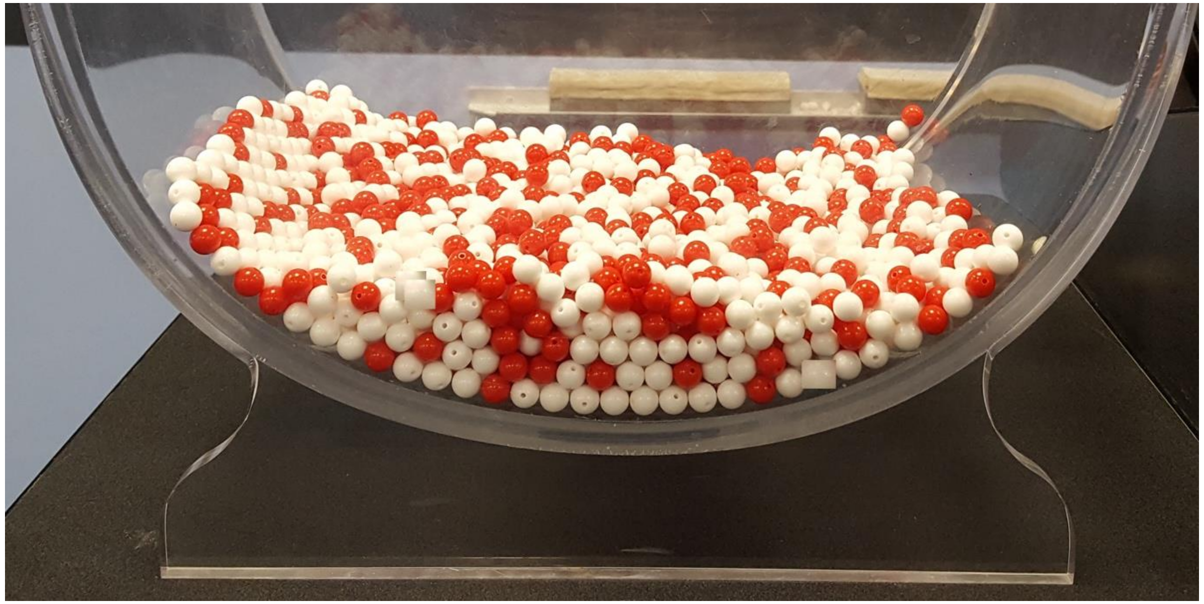
instats

Sampling

- Randomly selecting samples from a population
- Focuses on drawing multiple samples to study variability
- Key concept: population vs. sample

Population Proportion

- The ratio of subjects with a characteristic to the entire population
- Compute population proportion using `dplyr`
- **Tip**: Use `mean()` to find proportions of logical values



Sampling Variation

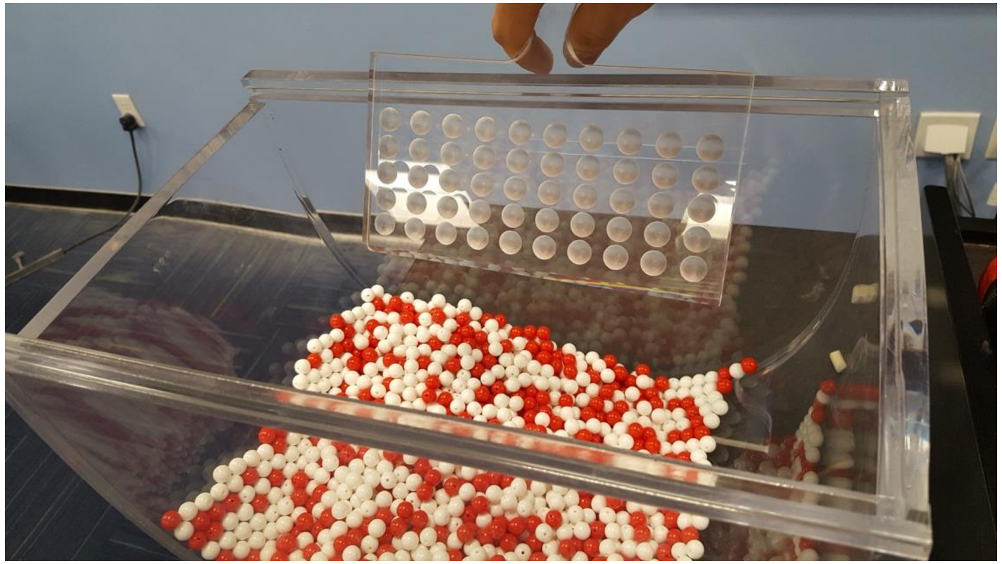
- Different samples produce different statistics due to random selection
- Key concept: Sampling with replacement and random mixing
- Explore sampling distribution with multiple replicates
- **Tip**: Histograms visualize the variation across samples

Central Limit Theorem (CLT)

- States that sample means will be approximately normally distributed
- Applies even if the population distribution isn't normal
- Key concept: The larger the sample, the better the normal approximation
- **Tip**: Look for bell-shaped histograms to confirm CLT

Sample Proportion

- The proportion of a particular outcome in a sample
- Important to understand sampling distribution of proportions
- **Tip**: Use `rep_slice_sample()` for repeated sampling



instats

Standard Error

- A measure of how much sample statistics (e.g., sample means or proportions) vary from the population parameter
- Decreases with increasing sample size
- **Tip**: The smaller the SE, the more precise the sample statistic

Virtual Sampling with Different Sample Sizes

- Study how the sampling distribution changes with varying sample sizes
- Smaller samples lead to greater variability
- Larger samples reduce variability and give a better estimate of the population parameter
- **Tip**: Visualize how the width of the distribution narrows with larger sample sizes

Demo & Exercises

Q & A

STOP

Welcome to

instats

The Session Will Begin Shortly

START

Statistics in R with Tidyverse

Session 8: Estimation via Theory-based Methods

instats

Estimation Primer

- Use sample data to estimate population parameters
- Focus on estimating the population mean
- Confidence intervals provide a range of plausible values
- Interpretation is key: they offer insight into the reliability of estimates

Understanding Confidence Levels

- Confidence intervals estimate the population parameter
- Provide a range of values rather than a single point
- Interval width reflects the uncertainty of the estimate
- Interpretation: "95% confident the population mean is within this range"

Constructing Confidence Intervals

- Based on sample mean and standard error
- Wider intervals indicate more uncertainty
- Narrower intervals suggest more precise estimates
- Confidence level determines the certainty we have about the interval

CLT and Confidence Intervals

- Central Limit Theorem (CLT) helps estimate the population mean
- With large samples, sample means follow a normal distribution
- Use CLT to construct confidence intervals even if population distribution is unknown
- Ensures validity of intervals for large sample sizes

Confidence Level

- Common levels: 90%, 95%, 99%
- Higher confidence means a wider interval
- Lower confidence results in a narrower interval
- Trade-off between precision and certainty

instats

Standard Error and Interval Width

- Standard error depends on sample size
- Larger samples reduce standard error
- Reduced error means narrower intervals
- Smaller samples result in wider intervals, increasing uncertainty

Using Confidence Intervals

- Useful for decision-making and interpreting data
- Practical for comparing groups or estimating trends
- Avoids over-reliance on single values (point estimates)
- Strengthens conclusions with a clear range

Interpreting Confidence Intervals

- Example: "We are 95% confident the mean lies between X and Y"
- Understand it’s about the method, not a 95% probability
- A narrower interval gives more insight, but requires larger sample sizes
- Consider context and trade-offs when interpreting

Demo & Exercises

Q & A

STOP

Welcome to

instats

The Session Will Begin Shortly

START

Statistics in R with Tidyverse

Session 9: Estimation via Bootstrapping Methods

instats

Bootstrapping Fundamentals

- Introduced by Brad Efron in 1979
- Bootstrap samples created by resampling with replacement
- Used to estimate population parameters like the mean
- Allows constructing confidence intervals, hypothesis testing, and more

The Bootstrap Method

- Random sample of size *n* taken from the population
- Bootstrap samples are created by resampling with replacement
- Some values appear multiple times, while others may not appear at all
- Repeat this process to generate many bootstrap samples

Why Use Bootstrapping?

- Allows estimation without knowing population distribution
- Provides insights even when traditional methods fail
- Works well with modern computing power
- Helps construct confidence intervals using resampled data

instats

Bootstrap Sample Mean Distribution

- Create many bootstrap samples, calculate sample means
- Distribution of these means forms the **bootstrap distribution**
- Resampling captures sampling variability
- Useful for constructing confidence intervals

Bootstrap Confidence Intervals

- Percentile method: Use the middle 95% of the bootstrap sample means
- Standard-error method: Use the bootstrap distribution standard error

for confidence interval width

- Resampling allows constructing accurate intervals even with limited data

Advantages of Bootstrapping

- Does not require assumptions about population distribution
- Flexible for various statistical estimations
- Can be applied to other parameters like medians or variances
- Widely used in fields with limited sample sizes

Demo & Exercises

Q & A

STOP